2021 IEEE International Conference on Smart Internet of Things (SmartIoT) | 978-1-6654-4511-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/SmartIoT52359.2021.00025

信頼度と特徴量の最適化に基づく単眼擬似LiDAR3D物体検出法

Jianlong Zhang
School of Electronic Engineering
Xidian University
Xi'an, China
jlzhang@mail.xidian.edu.cn

Guangzu Fang
School of Electronic Engineering
Xidian University
Xi'an, China
gzfang@stu.xidian.edu.cn

Xinyu Guo
School of Electronic Engineering
Xidian University
Xi'an, China
gxy8696@126.com

Yang Zhou
The Ministry of water resources of China
Beijing, China
zhy@mwr.gov.cn

Bin Wang 西安大学電子工程学院、中国 bwang@xidian.edu.cn

Ji Li
The Goldenwater Information Technology Co. Ltd.
Beijing, China
liji@goldenwater.com.cn

陳陳国家西安大学統合サービスネットワーク重点実験室、中国 cc2000@mail.xidian.edu.cn

概要-単眼Pseudo-LiDARに基づく3次元物体検出は、自律走行においてマルチセンサソリューションと比較してコスト効率が高く、近年大きな進歩を遂げている。しかし、これらの方法には、(1)Pseudo-LiDARの背景点の深度推定誤差が物体点の推定誤差よりも大きいため、物体検出性能が劣化する、という欠点がある。(2)既存の点群検出ネットワークは、Pseudo-LiDARの局所的な特徴量と大域的な相関を効果的に捉えることが困難であり、点群の特徴表現能力が弱い。このため、信頼度と特徴量の最適化に基づく擬似LiDARの3次元物体検出法を提案する。本手法では、まず点群信頼度最適化手法の階層構造を提案し、信頼度分布に従って背景点をフィルタリングするリサンプリングを行い、PseudoLiDARオブジェクト点群の有意性を向上させる。次に、DGCNNを用いてキーポイントの局所的な特徴を取得し、点群の特徴表現能力を向上させるために、キーポイントの大域的な関連性を捉える点変換器構造を持つ階層的な特徴抽出モジュールを設計する。一般的なKITTIベンチマークにおいて、我々のアプローチは主流手法よりも優れた性能を達成した。

キーワード-単眼3D物体検出、擬似LiDAR、信頼度最適化、階層的特徴抽出

I. Introduction

近年、3次元物体検出は、ロボットや自律走行アプリケーションなど多くのアプリケーションで重要な技術であり、物体をより現実的に記述する方法である。現在、3Dシーン情報を得るための関連手法は、自律走行ソリューションにおけるLiDARベースの手法、画像ベースの手法、複数のセンサーベースの手法に分類される。LiDARセンサーによって生成される点群の品質は、レーザースキャニングの特性、例えばレーザービームの数や回転速度に大きく依存する[1]。LiDARセンサーは、材料と技術の制約から高価であり、マルチセンサーソリューションのコストが高い。そのため、多くの研究が安価なカメラに置き換え始めた[1]。

現在、単眼カメラ[2]、ステレオカメラ、深度カメラ[3]が画像ベースの3D検出における主なセンサである。単眼画像は奥行き情報の不足により制限されており、他の2つの方式と比較して性能が劣る。しかし、成熟した技術、低コスト、工業化された生産に適したなどの利点があるため、依然として大きな注目を集めている。単眼画像に基づく3次元物体検出は大きな課題であるため、MonoGRNet[4]やM3D-RPN[5]では、2次元物体検出を画像から3次元物体検出空間に拡張しようとする手法もある。これらの方法は一定の結果を達成しているが、そのスケーラビリティは満足できるものには程遠い。

近年、Pseudo-LiDARに基づく3次元物体検出法が提案されている。本手法は、成熟した奥行き推定ネットワークを用いて、単眼画像から単眼奥行きマップを生成し、それを3次元点群に変換してシーン内の空間情報を表現し、LiDARベースの手法を利用して3次元バウンディングボックスを取得する。Pseudo-lidar[6]は、単眼奥行き推定法を利用してPseudoLiDAR点群を生成し、Frustum PointNet 3D検出フレームワークを用いて3Dバウンディングボックスを予測する。AM3Dは、点群データを生成するために推定深度を使用し、メッセージパッシングをガイドするために注意メカニズムを利用し、最終的に3Dオブジェクト検出性能に大幅な改善をもたらした[7]。しかし、[6,7]は単眼奥行き推定によって生成されるPseudo-LiDARに大きく依存しており、生成される点群の品質は最終的な検出性能に直接影響する。深度推定誤差の影響により、これらの手法の検出性能は悪く、改善の余地が大きい。

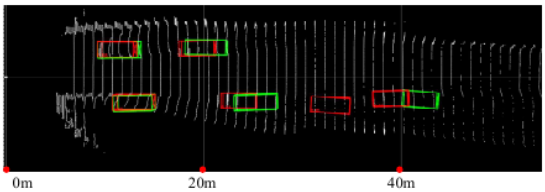

図1. KITTI[8]の3次元検出結果例。

図1は、KITTI3D[8]におけるPseudo-LiDARの「車」検証の鳥瞰図の典型的な結果である。グランドトゥルースのボックスは緑、予測されたボックスは赤で表示されている。物体の距離が長くなるにつれて、物体の検出性能が順次低下することが明らかである。

978-1-6654-4511-5/21/$31.00 ©2021 IEEE
DOI 10.1109/SmartIoT52359.2021.00025

これは、遠方の背景点群の深度推定誤差に起因するものである。さらに、LiDARベースの検出ネットワークは、例えば、優れたグローバル相関性能を享受するTransformer [9]を導入するなど、点群特徴の優れたオーバーエクストラクション手法の可能性をまだ持っている。

Pseudo-LiDARの信頼度と特徴量の最適化に基づく単眼3次元物体検出法を提案する。本論文の主な貢献は以下のようにまとめられる：(1) 単眼3D検出の検出ネットワークとしてPV-RCNN[10]を導入し、単眼3D検出のベンチマークを改善する。(2) 点群信頼度最適化手法の階層構造を提案し、信頼度分布に応じた物体点群の有意性を向上させ、奥行き推定誤差の影響を低減させる。(3) 局所特徴量と大域特徴量の両方を効果的に考慮し、点群の特徴表現能力を向上させる階層的特徴抽出モジュールを設計する。

本稿の残りの部分は以下のように構成されている。セクションIIでは、提案手法の原理と実装を紹介する。セクション III では、実験的な評価と分析を行う。最後に、セクションIVで我々の研究を締めくくる。

II. Approach

提案する単眼3D検出フレームワークは、図2に示すように、主に(1)Pseudo-LiDARの生成と最適化、(2)Pseudo-LiDARの物体検出の2つの部分から構成される。生成と最適化モジュールでは、まず単眼奥行き推定ネットワークに基づいて3次元視覚点群を生成し、次にカメラ座標系からLiDAR座標系に視覚点群をマッピングし、最後に信頼度最適化法を採用して物体の顕著性を向上させる。Pseudo-LiDAR物体検出では、PV-RCNNをバックボーンネットとして使用し、DGCNN[11]とPoint Transformer[12]を導入して、それぞれ点群の局所特徴と大域特徴を得るための階層的特徴抽出モジュールを構築する。

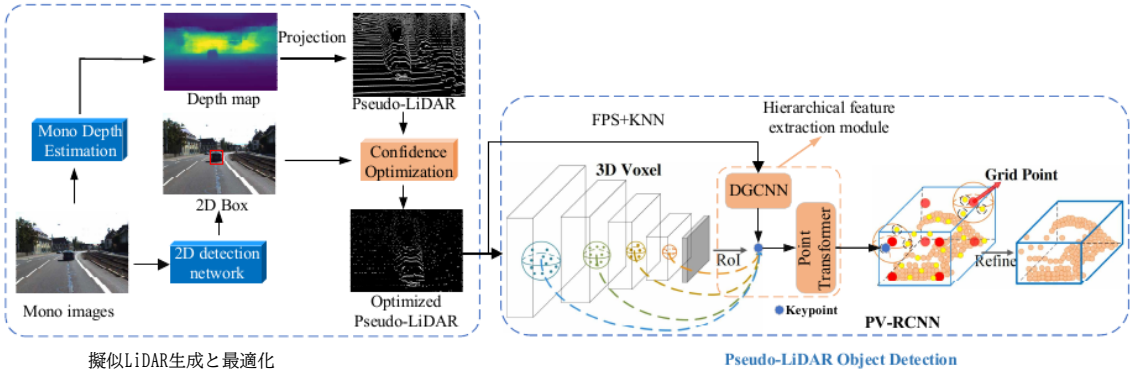

図2. 提案する単眼3D検出フレームワークの全体アーキテクチャ。(FPS: 最遠点サンプリング、KNN: K-最近傍)。

A. 擬似LiDAR生成

まず、単眼画像の奥行き推定ネットワークとしてDORN[13]を用い、このネットワークは順序回帰モデルを用いて長距離奥行き推定の精度を向上させる。

そこで、投影変換により3次元視覚点群にランダムに到達する、点群信頼度最適化のための階層的手法を提案する。最後に、図3に示すように、視覚的な点群を3次元座標変換によりカメラ座標からLiDAR座標に変換し、擬似LiDARを得る。

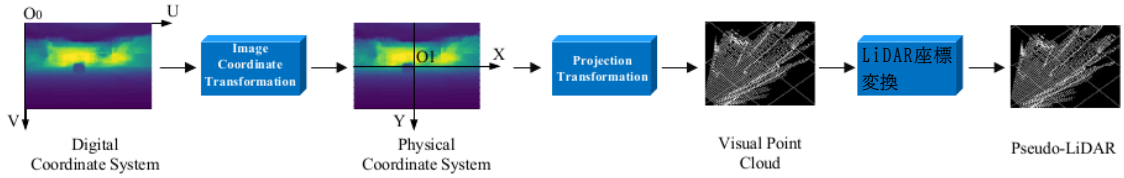

図3. 擬似LiDAR生成のための詳細な構造設計。

B. 信頼度の最適化

そして、深度マップをマッピングする単眼深度推定アルゴリズムには、以下の2つの欠陥がある。すなわち、1)物体車両の端の深度値が背景の影響を受け、その結果、物体と背景の間の遷移領域における深度推定誤差が増加する、2)背景点の深度推定誤差が物体点のそれよりもはるかに大きく、距離とともに非線形に増加する、である。

は、局所的および大域的な信頼度に従ってPseudo-LiDARを再サンプリングし、大きな奥行き推定誤差を持つ背景点をフィルタリングし、物体点の有意性を向上させる。

図4に示すように、信頼度最適化手法は主に以下の部分から構成される:1)2D画像中の物体位置を特定し、物体と背景を区別するための2D検出モジュール、2)物体と背景の間の遷移部分の点の信頼度を低減するための局所信頼度モジュール(LCM)、

3)長距離点の信頼度を低減するためのグローバル信頼度モジュール(GCM)、4)信頼度リサンプリングモジュールfまたは、奥行き推定誤差が大きい背景点をフィルタリングする。

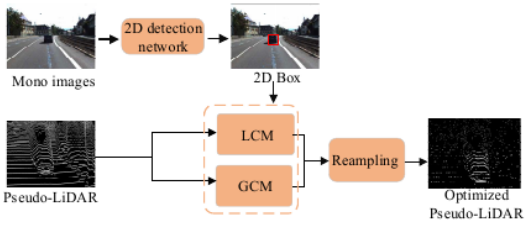

図4. 信頼度最適化手法

1) 2次元検出ネットワーク: カスケードR-CNN[14]は検出モデルであり、2次元検出ネットワークとする。従来のR-CNNネットワークと比較して、カスケード方式で複数回物体を検出し、サンプル数を変えることなく高性能な2次元検出ネットワークを学習する。ここで注目すべきは、カスケードネットワークに起因するオーバーフィッティングの問題は、再検出によって軽減されることである。

2) 局所信頼度モジュール(LCM): 検出バウンディングボックスの局所信頼度が2次元ガウス分布に従うと仮定し、2次元検出バウンディングボックスの中心をアンカー点とし、点群とアンカー点間の距離に応じて局所信頼度を計算する。この方法は、車両エッジ点群の信頼性を低下させ、検出に対する奥行き推定誤差の影響を弱める。

擬似LiDAR点群における点の空間座標をp(x,y,z)とし、2次元画像座標系上の投影点をa(u,v)と定義する。投影点aが2次元検出バウンディングボックスbの内側にある場合、投影点aのbに対するガウス分布の重みは$\alpha_{(a,b)}$として定義される。

$$\alpha_{(a,b)} = \begin{cases} s(u,v) & y(a,b)=1 \\ 0 & y(a,b)=0 \end{cases}, \tag{1}$$

ここで、y(a,b)=1は、aが検出バウンディングボックスbの内側に位置することを意味する。s(u,v)の計算式は次の通りである。

$$s(u,v) = \frac{1}{2\pi\sigma^2} e^{-\frac{(u-u_c)^2+[(v-v_c)*\frac{w}{h}]^2}{2\sigma^2}}, \tag{2}$$

ここで、wとhはそれぞれ検出バウンディングボックスbの画素幅と高さ、減衰率パラメータσはw/5、(u_c, v_c)は検出バウンディングボックスbの中心座標である。

オブジェクト車両がオクルージョンされている場合、2D検出のバウンディングボックスが重なり合うことがあり、1つの点群pが複数の検出により複数のバウンディングボックスに変換されることがある。このとき、最大信頼度重みは、1つの点群pの重みを計算するために使用される。$B=\{b_1, b_2, \cdots, b_m\}$を画像シーンにおける投影点aのすべての検出バウンディングボックスを含む集合とすると、投影点aの重み$\alpha_{(a,B)}$は

$$\alpha_{(a,\mathbf{B})} = \max(\alpha_{(a,b)}), b \in \mathbf{B}. \tag{3}$$

とすると、擬似LiDARの局所信頼度S_{Loc}(p)は

$$S_{Loc}(p) \max(f_{norm}(D_{(a,B)}), [_D), \tag{4}$$

ここで、f_{norm}はPseudo-LiDAR点群セットのシーン重み正規化関数であり、ξ_αは検出バウンディングボックスの外の背景点が完全にフィルタリングされないようにするための局所背景閾値である。

3) グローバル信頼度モジュール(GCM): 単眼画像では奥行き事前情報がないため、奥行き推定誤差はシーンの距離が長くなるにつれて非線形に増加する。このことを考慮し、信頼度減衰率R_γを設計することで、距離の増加とともに大域的な信頼度分布が減少し、長距離奥行き推定誤差の検出への影響が弱まる。

異なるシーンにおけるPseudo-LiDARの深度分布は全く異なるため、現在のシーンのPseudo-LiDARはQ(p)と定義される。まず、グローバルバランスパラメータλ_βを導入し、点群深度分布に従って信頼度減衰率R_γを次のように計算する。

$$R_\gamma = \frac{1}{\lambda_\beta f_E(Q) + f_D(Q)}, \tag{5}$$

ここで、f_E(Q)とf_D(Q)はそれぞれPseudo-LiDARの深度平均と深度分散である。

$$f_E(Q) = \frac{\sum_{p\in Q} d_p}{|Q|} \quad , \quad f_D(Q) = \sqrt{\frac{\sum_{p\in Q}[d_p - f_E(Q)]}{|Q|}}, \tag{6}$$

ここで、d_pは点群pの深度値である。

すると、Pseudo-LiDAR点群のグローバル信頼度計算式は

$$S_{Global}(p) = \max(1 - R_\gamma d_p, \xi_\beta), \tag{7}$$

ここでξ_βはグローバルバックグラウンド閾値である。

式(7)は、S_{Global}(p)が点群深度の増加とともにR_γの割合で減衰することを示している。背景閾値ξ_βは、遠距離の点群が完全にフィルタリングされないことを保証する。

最終的な擬似LiDAR点群信頼度S(p)は、S_{Global}(p)をS_{Loc}(p)に重み付けすることで生成される、

$$S(p) = S_{Loc}(p) \cdot S_{Global}(p). \tag{8}$$

4) 信頼度の再サンプリング: 生の擬似LiDAR集合Q_{raw}の各点は、対応する信頼度に従って再サンプリングされ、再サンプリングされた点群集合はQ_{re}として定義される。

$$Q_{re} = \{p \mid S(p) > rand(0,1), p \in Q_{raw}\}. \quad (9)$$

Pseudo-LiDARの信頼度リサンプリング前後の結果をそれぞれ図5の左と右に示す。背景点と物体端点が疎になり、物体点群の顕著性も向上していることがわかる。

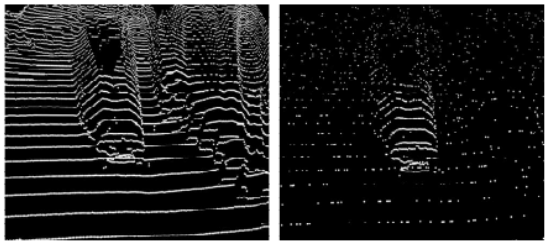

図5. 信頼度リサンプリング。(生の擬似LiDAR(左))。リサンプリング擬似LiDAR(右))

C. 階層的特徴抽出構造に基づく擬似LiDAR物体検出ネットワーク

PV-RCNN[10]に基づく擬似LiDAR物体検出ネットワークを提案する。PV-RCNNはキーポイントの特徴を抽出することで、3Dボックス提案を洗練する。つまり、キーポイント特徴の品質は最終的な検出性能に直接影響する。各キーポイントは最遠点サンプリング(FPS)によって得られるが、スパースサンプリング処理により特徴量が失われる。そして、各キーポイントの特徴は、点群の局所近傍特徴と大域的関連性を考慮しないPointNetベースの集合抽象化によって抽出される。以上のことから、生のPV-RCNNネットワークはPseudo-LiDARの特徴を効果的に抽出できなかったことがわかる。そこで、PV-RCNN物体検出ネットワークに基づく階層的な特徴抽出構造を提案し、図6に示す特徴抽出モジュールは、主にDGCNN[11]とPoint Transformer[12]構造で構成され、それぞれキーポイントの局所的特徴と大域的特徴を得る。

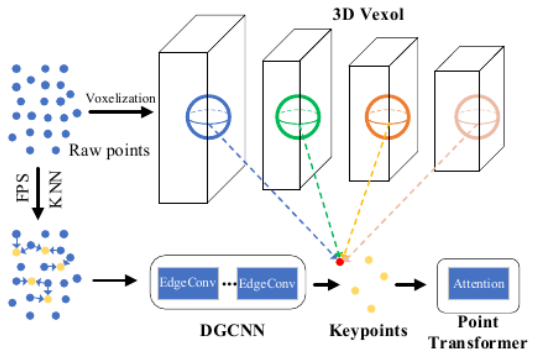

図6. 階層的特徴抽出モジュール

1) 局所特徴抽出: まず、KNN(Knearest neighbor)によりキーポイントの局所近傍構造を構築し、キーポイントの特徴をキーポイントに集約することで、サンプリングキーポイントに生点の特徴情報が多く含まれるようにする。ここで、$\chi=\{x_i\}_i$ はキーポイントの特徴ベクトル集合である。

であり、キーポイントの局所近傍特徴は、EdgeConvをカスケード接続することで高次元局所特徴空間にマッピングされる。

EdgeConvの定義は以下の通りである。$e_{ij} = h_\Theta(x_i, x_j - x_i)$ はエッジ特徴量であり、h_Θ は学習可能なパラメータの集合 Δ [11] でパラメータ化されたパラメトリック非線形関数である。Aは最大プーリング集約演算であり、エッジ特徴の最大プーリング演算は、式(10)に示すように、EdgeConvと呼ばれる:

$$x_i' = \underset{j:(i,j)\in\varepsilon}{\mathrm{A}} h_\Theta(x_i, x_j - x_i), \quad (10)$$

ここで{j : (i, j)ε}を中心画素x_iの周りのパッチとする(図7参照)。EdgeConvの出力は、図7に示すように、すべてのエッジ特徴を集約して計算される。

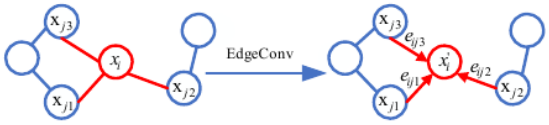

図7. EdgeConv操作の可視化。

2) 大域的特徴抽出: 本論文では、PV-RCNNネットワークがキーポイントの大域的な相関特性を捉える能力を向上させ、物体形状に対するネットワークの知覚を向上させるために、Point Transformer構造を導入する。Transformer [9]は近年、自然言語処理における傑出した貢献である。主に自己注意メカニズムを用いて、シーケンスの長期的な相関を捉える。最近の研究では、変形変換器による点群セグメンテーションの分野に拡張されている[12]。Point Transformerはベクトル注意を使用し、そのベクトル注意の重みは複数の特徴チャンネルを調整することができる。点群は本質的にメトリック空間に埋め込まれた不規則な集合であるため、自己注意は点群に自然に適用できる。

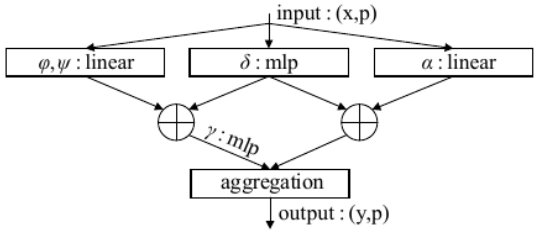

図8. 点変換層

Point Transformerの注目度表現は

$$y_i = \sum_{x_j \in \chi(i)} \rho\left(\gamma\left(\varphi(x_i) - \psi(x_j) + \delta\right)\right) \odot \left(\alpha(x_j) + \delta\right), \quad (11)$$

y_i はキーポイントの出力特徴量である。ρは正規化関数(例えばソフトマックス)である。γはMLPのような写像関数である。φ, ψ, α はキーポイントの特徴変換である。δは位置エンコーディング関数である。Point Transformer層の構造を図8に示す[12]。

点変換器は多くの GPU メモリを消費するため、キーポイントをダウンサンプリングするために FPS を選択し、キーポイントと大域的特徴の間の長距離相互作用を得るために、キーポイントを点変換層に出力する。

III. EXPERIMENTS

A. 実験の詳細

1) データセット KITTI[8]は3D自動運転で最も人気のあるデータセットで、学習とテストに使用され、7481個の学習サンプルと7518個のテストサンプルを持つ。我々のネットワークの性能をテストするために、トレーニングサンプルはトレーニングセット(3712サンプル)と検証セット(3769サンプル)に分けられ、これらは我々のネットワークモデルのトレーニングとテストに使用される。

2) パラメータ設定：局所信頼度の計算では、局所背景閾値ξ_αを0.2とする。大域的信頼度の計算では、大域的バランスパラメータλ_βを1.5、大域的背景閾値ξ_βを0.2とする。サンプリング点数は2048点、最近傍点数Kは16点である。

3) 学習：深層学習フレームワークをPytorchでコーディングする。コサインアニーリング戦略に従って減衰するAdamオプティマイザを使用する。GTX 2080Ti GPUでネットワークを学習させた。

4) メトリック KITTIベンチマークが提供する公式評価ツールを用いて、Intersection of Union (IOU)の閾値0.7の下での3次元平均精度を計算し、11個の補間点(AP_{11})を用いて「Car」カテゴリのAP値を計算する。ベンチマークは、バウンディングボックスの高さとオクルージョンレベルに応じて、「車」カテゴリを「簡単」「中程度」「難しい」の3つのケースに分割していることに注意。

B. 性能比較

本手法と主流の4つの単眼3D検出手法MonoGRNet、M3D-RPN、pseudo-LiDAR、AM3Dの3D "Car "物体検出結果を表1に示す。本手法の性能はAP_{11}とIOUの評価指数が0.7より大きい場合、他の単眼検出法より優れている。3つのケース(easy, moderate, hard)において、最良のAM3Dと比較して、我々の手法のAP_{11}値はそれぞれ6.92%、5.76%、5.39%増加した。

TABLE I. KITTI 3D "CAR "バリデーションセット

Method	*Easy(%)*	*Moderate(%)*	*Hard(%)*
MonoGRNet [4]	13.88	10.19	7.62
M3D-RPN [5]	20.27	17.06	15.21
Pseudo-lidar [6]	28.20	18.50	16.40
AM3D [7]	32.23	21.09	17.26
Our method	**39.15**	**26.85**	**22.65**

C. Ablation experiment

すなわち、(1)我々の手法から階層的特徴抽出モジュール(HFEM)と信頼度最適化モジュール(COM)を削除し、(2)我々の手法から階層的特徴抽出モジュール(HFEM)を削除した。アブレーション解析の結果を表2に示す。最も性能の良いモデルは、COMとHFEMを組み合わせたモデル、すなわち提案手法である。表2に示すように、COMとHFEMの両方が単眼3D検出の性能を向上させることができる。上記3つの実験の3D検出結果を図9に、グランドトゥルースのボックスを緑、予測されたボックスを赤で示す。COMを用いたネットワークでは、誤検出バウンディングボックスの数が減少しており、これは奥行き推定誤差の影響を低減し、検出性能を向上させることができることを示している。HFEMを追加した後、長距離オブジェクトの検出精度が向上し、全体的な検出性能も向上している。

TABLE II. KITTI 3D "CAR "バリデーションセット

Method	*Easy(%)*	*Moderate(%)*	*Hard(%)*
(1) Remove HFEM and COM	35.24	24.55	20.75
(2) Remove HFEM	36.96	25.84	21.13
(3)The full framework	**39.15**	**26.85**	**22.65**

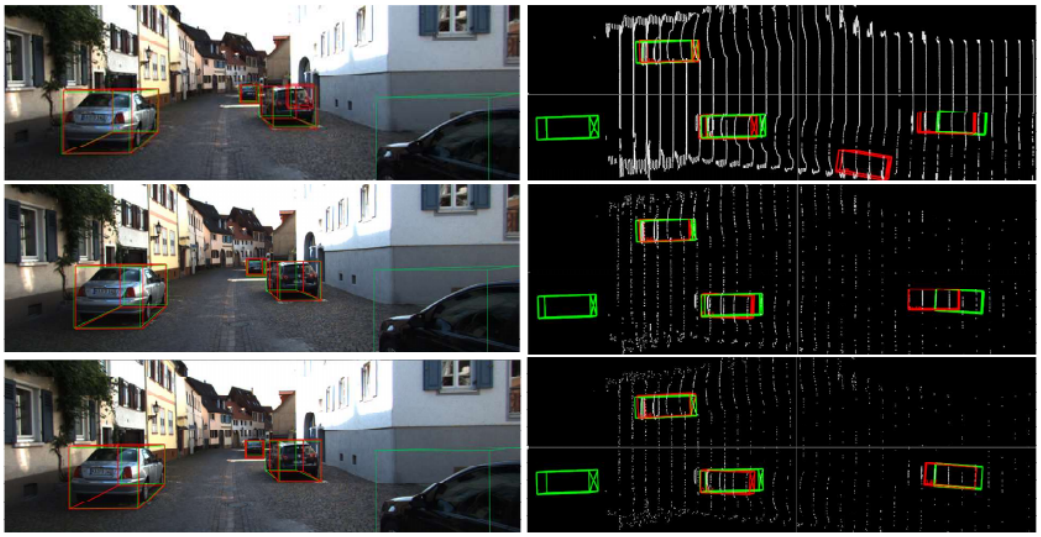

Fig. 9. Qualitative comparisons of 3D detection results. (*The 3D detection results of RGB image* (*Left*), *the 3D detection results of Pseudo-LiDAR* (*Right*). *Removing HFEM and COM* (*top*), *removing HFEM* (*middle*), *and the full framework* (*bottom*).)

IV. Conclusion

本論文では、信頼度と特徴量に最適化されたPseudo-LiDARに基づく単眼3D検出法を提案する。物体検出の基幹ネットワークとしてPV-RCNNを用いることで、Pseudo-LiDAR物体の検出能力を向上させる。信頼度最適化により、オブジェクト点群の重要度が向上し、階層的な特徴抽出構造により、キーポイントの特徴表現能力が向上する。シミュレーション実験の結果、提案手法は、主流の単眼3D検出手法と比較して、KITTI上の「車」物体の3つのケースにおいて明らかな優位性を持つことが示された。さらに、より正確な深度推定マップがあれば、我々の検出ネットワークの検出性能はより良くなる可能性がある。

Acknowledgment

本研究は、中国航空科学基金(No.2018ZC81001)、中国国家自然科学基金(No.61971331)、中国国家重点研究開発プログラム(2018YFE0126000)、中国国家自然科学基金(62072360、61902292、62001357、62072359、62072355)、陝西省重点研究開発計画(2021ZDLGY02-09、2019ZDLGY13-07、2019ZDLGY13-04、2020JQ-844)、組込みシステム・サービスコンピューティング重点実験室(同済大学)(ESSCKF2019-05)、文部科学省西安科学技術計画(20RGZN0005)、西安モバイルエッジコンピューティング・セキュリティ重点実験室(201805052ZD3CG36)。

References

[1] Vianney, J. M. Uwabeza, S. Aich, and B. Liu, "RefinedMPL: Refined Monocular PseudoLiDAR for 3D Object Detection in Autonomous Driving," arXiv preprint arXiv: 1911.09712, 2019.

[2] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 270-279.

[3] K. Yamaguchi, D. McAllester, and R. Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in European Conference on Computer Vision. Springer, Cham, 2014, pp. 756-771.

[4] Z. Qin, J. Wang, and Y. Lu, "Monogrnet: 単眼3次元物体定位のための幾何学的推論ネットワーク", AAAI人工知能会議論文集, vol.33, no.01, pp.88518858.

[5] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9287-9296.

[6] Y. Wang, W. L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8445-8453.

[7] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan, "Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6851-6860.

[8] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354-3361.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," arXiv preprint arXiv:1706.03762, 2017.

[10] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, and X. Wang, "PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10529-10538.

[11] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," ACM Transactions on Graphics (TOG), 2019, 38(5): 1-12.

[12] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point Transformer," arXiv preprint arXiv: 2012.09164, 2020.

[13] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2002-2011.

[14] Cai, Z. , and N. Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154-6162.